Prepared for:

Federal Communications Commission

CMS Alliance to Modernize Healthcare
Federally Funded Research and Development Center

Contract No. HHSM-5000-2012-000081
Task Order No. FCC15D0002

Internet Protocol Caption Telephone Service (IP CTS) Devices: Summary of Phase 1 Activities

Final

Version 2.0

July 24, 2017

Executive Summary

The Federal Communications Commission (FCC) engaged The MITRE Corporation (MITRE) through the Centers for Medicare & Medicaid Services (CMS)-sponsored CMS Alliance to Modernize Healthcare (CAMH) Federally Funded Research and Development Center (FFRDC) to independently assess the quality metrics and associated usability of Internet Protocol Caption Telephone Service (IP CTS) devices and services, as well as alternative technologies that could be used in place of IP CTS. Given the substantial growth of IP CTS usage in the last two years and advances in technology, the FCC is interested in understanding whether new technologies or processes can provide improved IP CTS service while continuing to ensure that IP CTS services are appropriately available to individuals who need assistance to obtain equivalent access to telephony services.

Toward these ends, MITRE conducted a usability survey, a usability study, and independent device testing to identify characteristics of IP CTS performance and use. These activities encompassed a broad spectrum of IP CTS users (540 survey respondents, 20 usability test subjects) and all IP CTS providers' devices.

What We Found

Initial Survey and Usability Study Results

The survey and usability study results in Phase 1 indicate the need for more research and testing to fully grasp the usability and needs of the community. Certain results from the 540 respondents, however, are clear:

1. Users are dissatisfied with delay (41% of respondents) and accuracy (32% of respondents) of transcription services.
2. Some device features can be adjusted to better accommodate the hard of hearing community. Specifically, audio indicators are not necessarily helpful as observed during assessments. Other indicators for button press (haptic), dial tone available, and silence on the line would be helpful.
3. Speakerphone or bi-aural headset capability is helpful. Some members of the hard of hearing community can comprehend speech better if they can use both ears as noted from user comments during the assessments.

Device Testing Results

1. Automated Speech to Text (STT) tools can provide much lower transcription delay (deemed better). In all but one case, the STT tools provided accuracy at least as good as the worst of the IP CTS providers. In two cases, automated STTs provided better accuracy than any of the IP CTS providers.
2. Some providers optimize for accuracy over transcription delay and others for delay over accuracy. The current survey and usability results do not conclusively establish whether accuracy or delay is more important to users or what constitutes "tolerable levels" for either measure.

3. Because "accuracy" is not well defined, a rigorous definition is important. MITRE is working with the Telecommunications Relay Service Center of Excellence and other resources to identify the important aspects of accuracy to establish a rigorous, IP CTS-specific definition.

Table ES-1 summarizes caption delay and accuracy for test calls from device testing and the usability assessment by provider.[1]

Table ES-1. Average Transcription Delay and Accuracy by Provider

Provider	Average Transcription Delay (Seconds)	Average Accuracy (Percentage)
Provider 1	15.8	88.3
Provider 2	7.3	84.5
Provider 3	4.1	82.8
Provider 4	14.6	88.7
STT-1	2.2	83.0
STT-2	2.1	75.6

Recommendations

MITRE recommends that the FFC should:

1. Research the possibility of using fully automated Speech to Text services in place of existing IP CTS services. Continue usability testing to determine if automated STT system prototypes can provide (a) similar levels of usability as experienced today by IP CTS users, (b) an overall satisfactory calling experience similar to what is currently available, and (c) are viable options/alternatives to IP CTS services.
2. Continue working with the hard of hearing community to identify the key measures and metrics for telephony captioning comprehension and usability.
3. Establish a quality assurance group to test transcription quality.
4. Request that IP CTS providers include speakerphone and headphone capabilities in devices.
5. Request that IP CTS providers include visual or tactile feedback for all functions that currently use audio feedback (e.g., dial tone or silence indicators, button-pressed indicators).

[1] MITRE included two unidentified automated Speech to Text providers (STT-1 and STT-2) that offer freely available automated SST tools over the Internet.

Record of Changes

Version	Date	Author / Owner	Description of Change
1.0	November 18, 2015	MITRE	Final, Version 1.0 for delivery to sponsor
2.0	July 24, 2017	MITRE	Final Version 2.0, for sponsor release to public

Table of Contents

List of Figures

List of Tables

1. Introduction

Hearing loss is an invisible but significant barrier in daily life, including telephone conversations. According to a 2012 study by the United States Census Bureau,[2] 7.6 million (3.1%) of the U.S. population experienced a hearing difficulty—defined as experiencing deafness or having difficulty hearing a normal conversation, even when wearing a hearing aid. Severe hearing loss affects 2.2 million or 0.8% of the population. For those 65 and older, 4.2 million (10.8%) experienced a hearing difficulty, including 1.7 million (4.3%) who reported a severe difficulty hearing. As the U.S. population ages, the number of individuals with hearing or vision loss is projected to increase significantly.

Internet Protocol Caption Telephone Services (IP CTS) is a telecommunications relay service for an individual who can speak, but who has difficulty hearing over the telephone. An individual can use a telephone and an Internet Protocol-enabled device to listen to the other party and simultaneously read captions of the other party's words.

The Federal Communications Commission (FCC) requested The MITRE Corporation (MITRE), as operator of the Centers for Medicare & Medicaid Services (CMS) Alliance to Modernize Healthcare (CAMH) Federally Funded Research and Development Center (FFRDC), to independently assess the quality and usability of IP CTS devices and services as well as alternative technologies that could be used to augment current IP CTS technologies.

1.1 Document Purpose and Scope

This document consolidates the results of the IP CTS survey, usability testing, and device testing activities during Phase 1. MITRE conducted these three activities independently. The following documents provide additional details on the testing processes and results:

- *Internet Protocol Caption Telephone Service (IP CTS) Devices: Usability Survey Report*
- *Internet Protocol Caption Telephone Service (IP CTS) Devices: Usability Testing Results*
- *Internet Protocol Caption Telephone Service (IP CTS) Devices: Baseline Test Results*

1.2 Study Goals and Assessment Objective for Phase 1

The objective of the independent assessment conducted in Phase 1 was to collect and provide the FCC with IP CTS user preference data and device characteristics that the FCC can use to inform policy changes regarding performance, quality, and contextually relevant standards for IP CTS service providers.

MITRE and the FCC intend that these activities will help the FCC and other organizations establish a standard methodology for measuring user preference on the quality of IP CTS device performance. In collaboration with the FCC, MITRE will define the study goals further and assess the objectives for Phase 2.

[2] M. Brault. (2012). "Americans with Disabilities: Household Economic Studies." U.S. Census Bureau. http://www.census.gov/prod/2012pubs/p70-131.pdf

1.3 Overall IP CTS Testing Approach

This study consists of two phases. The summary results presented in this document *address specific findings for Phase 1 only*. The phases of testing are as follows:

Phase 1 – Baseline of IP CTS

1. Internet Protocol Caption Telephone Service (IP CTS) Devices: Initial Baseline Test Results
    a. Establish appropriate performance measures and quality standards for IP CTS. The data includes capturing baseline technical measures with a focus on accuracy, latency, and completeness.
2. Baseline Usability Assessment of IP CTS with Users
    a. Utilize performance measures and quality standards from Part 1 and collect usability feedback, comparing objective performance measures of IP CTS and feedback from users.
3. Utility of the Service (Survey of Users)
    a. Attitudinal survey of the hard of hearing community regarding IP CTS use and areas for potential improvement with a focus on usability and user experience.
    b. Identify and categorize user demographics.

Phase 2 – Testing of Existing Alternative Technologies and Approaches, with three parts:

1. Identification and Testing of Commercially Available Alternatives to IP CTS Equipment
2. Usability Assessment (controlled User Testing)
3. Utility of Service Testing of Alternatives (Survey of Users)

MITRE's intention is to use the Framework to provide repeatable, consistent, and valid test results.

Phase 1 consisted of the following activities.

1.3.1 IP CTS Survey

MITRE partnered with Gallaudet University (GU) and Rochester Institute Technology (RIT) to develop, recruit participants for, conduct, analyze, review, and report on the IP CTS national attitudinal survey. The Gallaudet University Technology Access Program and the National Technical Institute for the Deaf (NTID) from RIT provided invaluable assistance in developing the survey, and recruiting participants.

The goal of the national survey was to understand (1) demographics, including severity of hearing loss; (2) when users require captioning services; (3) usability issues encountered; and (4) opportunities for service and device improvement.

1.3.2 Usability Study

MITRE conducted controlled user assessments to establish a baseline of usability metrics based on the assessment of IP CTS devices and services. The testing environment included four provider devices. Participants in the video-captured study included 20 hard of hearing users who completed the two (2)-hour usability assessment.

1.3.3 IP CTS Device Testing

MITRE conducted independent testing of IP CTS devices and services in a controlled environment. This baseline testing provides quantitative measures for key performance characteristics of the IP CTS service. The baseline testing does not provide pass/fail criteria or identify acceptable quality standards for the IP CTS service. These measures are used to support usability testing and, in the subsequent assessment phase, may be used to identify changes in IP CTS service quality.

2. Measure and Metric Definitions

The definitions in this section apply consistently across the user survey, usability study, and device testing results.

2.1 Device-related Metrics

MITRE used the following metrics to measure device and service-related quality:

1. **Time for a Communication Assistant (CA) to Connect**
    a. Definition: The elapsed time between a user's request for captions (e.g., by pressing the "Captions On" button during a call) and the display of text on the phone's screen, indicating the start of captioning services.
    b. How to measure: From the video recording of each call, MITRE analysts measure and record the elapsed time (in seconds) between the user's request for captions and the notification of CA assigned to the call. Generally, the notification includes a CA-specific number to indicate the call is being captioned. If this identifier is not displayed, MITRE will look for another textual notification of the captioning service.
2. **Accuracy**
    a. Definition: The percentage of words from the conversation (the IP CTS transcription) that are correctly transcribed on the Device under Test (DUT) screen.
    b. How to measure: Using the video recording of each call, MITRE analysts create a text file of the captions displayed on the DUT screen (the IP CTS transcription). MITRE uses the National Institute of Standards and Technology (NIST) SCLITE tool from the Speech Recognition Scoring Toolkit (SCTK)[3] to compare this file to a reference "true transcription". MITRE records the tool's scores and output.

    The SCLITE tool scores accuracy by comparing the total number of words transcribed correctly to the number of words in the reference file. Omitted words are reflected in both the accuracy and completeness metrics.

    There may be more than one correct transcription of spoken words to written words. To provide a consistent, realistic assessment of accuracy, MITRE has identified the following rules for assessing transcribed files:

    - Uppercase/Lowercase and punctuation are not considered in the accuracy calculations. All files are forced to uppercase. All punctuation is removed prior to assessment except for hyphens ("-"), underscores ("_"), and slashes ("/").
    - Contractions and expanded phrases are both valid ("that's" and "that is" are considered the same).

[3] http://www.itl.nist.gov/iad/mig/tools/

- Abbreviations that have spaces or periods between the letters are considered valid (“FCC”, “F C C”, and “F.C.C” are all considered the same).
- Numbers may be spelled out or numeric (“400”, “four hundred”, and “4 hundred” are all considered the same).
- Times may be represented as words or numbers (“8:30”, “8 30”, “830”, and “eight thirty” are all considered the same).
- Hyphenated words, non-hyphenated words, and words separated by underscores are all valid (“thank you”, “thank-you”, and “thank_you” are all considered the same).
- Universal Resource Locators (URL) that contain extra spaces or spell the words “slash” or “dot” are valid (“fcc.gov/smartdevice”, “fcc dot gov slash smart device”, and “fcc . gov / smartdevice” are all considered valid).
- Disfluencies (“ah”, “um”, “hmm”) may be omitted, but are not counted as errors if included.
- Singular instead of plural, and vice versa, will be counted as incorrect (“hour” is not the same as “hours”).
- Homophones will be counted as incorrect (“their”, “there”, and “they’re” are not considered the same; “Press for if” and “press 4 if” are not considered the same).
- Concatenated words are considered correct if the concatenated word has substantially the same meaning as the individual words (“video games” and “videogames” are considered the same, whereas “indecent” and “in decent” are not considered the same).
- Address abbreviations are valid. For example, “South 16th East” and “S. 16 E.” are considered the same.

3. Readability

a. Definition: The grade level (based on the U.S. education system) at which a user can understand text.

b. How to measure: MITRE analysts use the reference “true transcription” for each audio file to calculate grade-level readability and comprehension based on the Flesch-Kincaid reading ease formula score with an online tool. Readability is a function of the complexity of the audio file, not the transcription. A conversation with higher complexity should have more transcription errors.

4. Reading Ease

a. Definition: The Flesch reading-ease test uses a formula to rate the ease of readability wherein higher scores indicate material that is easier to read; lower numbers mark passages that are more difficult to read.

b. How to measure: MITRE analysts use the control for each audio file to calculate reading ease based on the Flesch-Kincaid reading ease formula score with an online tool (https://readability-score.com). Readability is a function of the complexity of the audio file, not the transcription.

i. Scores > 90 should be easily understood by an 11-year-old student.

ii. Scores 60 – 70 should be understood by a 13- to15-year-old student.

iii. Scores < 30 are best understood by university graduates.

5. **Caption Delay (Latency)**

   a. Definition: The time elapsed between hearing a voice on the caption phone and the display of captions on the phone's screen.

   b. How to measure: For each audio file, MITRE analysts identify eight words spaced uniformly throughout the recording. Based on the video recording, MITRE measures and records the time between when each word is heard and when it appears in the transcript. If the word does not appear in the transcript, MITRE records the time until the next word appears or, if there is a long gap in conversation before the next word would be expected to appear, MITRE records the time for the prior word to appear. MITRE recorded delay times with single-second resolution. All values less than one second were recorded as one second. MITRE rounded up delay values where the fractional part of the measurement was greater than "0.5" to the next second.

6. **Completeness**

   a. Definition: A measure of the words transcribed correctly or incorrectly as a percentage of the total words in the audio file.

   b. How to measure: MITRE analysts use the number of words in the "true transcription" and then compare the IP CTS transcription to determine the number of words not included in the output. Incorrectly transcribed words are considered "included". Completeness is calculated as the number of words transcribed divided by the number of words in the original audio file, expressed as a percentage. From the video recording of each call, MITRE creates a text file of the transcribed audio. MITRE uses the NIST SCLITE tool to compare the IP CTS transcription file to the "true transcription". MITRE records the SCLITE tool scores and output.

2.1.1 Usability Metrics

MITRE captured participants' feedback on the varying degrees of call quality to identify a relationship between levels of system performance and usability feedback.

After each participant completed a call or test case, MITRE queried the users on the following usability dimensions via 7-point Likert Scale statements:

- **Effectiveness –** The degree to which a user continuously maintains full context and meaning of the conversation.
- **Efficiency –** The degree to which a user is comfortable with the perceived tempo of the call.
- **Satisfaction –** The degree to which the user is satisfied with the calling experience, given the test case and use of transcript.

- **Overall Usability –** Computed average of the effectiveness, efficiency, and satisfaction metrics.

Note**:** For participants with severe hearing loss, the results do not include audio-focused questions. See Appendix D for Usability Questionnaire.

3. Device and Service Measurement

During device testing and usability study for the first phase, MITRE collected baseline performance metrics using the IP CTS devices listed in Table 1. The objective of this phase was to collect system performance and usability information on the proposed quality metrics presented in subsection 2.1. To assure accuracy, MITRE referenced and used existing NIST tools and standards for automatic speech recognition (ASR).[4]

MITRE collected baseline performance metrics using a controlled environment with pre-recorded audio for device testing and from test calls made during usability testing.

MITRE confirmed with the IP CTS providers that the setup and initialization of lab equipment was comparable to equipment fielded in users' homes and businesses. During preliminary interviews, Provider 1 indicated that the specific device type under test should have no impact on the captioning service: all devices use the same codec and underlying technology for communication with the captioning service.

MITRE also tested two automated Speech-To-Text (STT) engines that are freely available over the Internet. These two Internet STTs are indicated as STT-1 and STT-2 in the results summaries.[5]

3.1 Device and Service Test Results

The following subsections summarize device and service quality metrics for each conversation tested. MITRE used seven conversation types for call testing. Appendix A contains the reference text for each conversation. Appendix B contains more detailed quality metrics for each conversation used in testing.

3.1.1 Time for a CA to Connect

Time for a CA to Connect was the one area where testing revealed differences between providers. As Table 1 shows, the average time for an IP CTS CA to connect varied between 3.6 and 5.9 seconds, with Provider 2 demonstrating longer connect times than other providers. Time to connect was not applicable for automated STTs because these services do not rely on telephony connections and are invoked differently than the IP CTS providers. The average time to connect was based on 30 test calls per provider from device testing.

Table 1. Average Time to Connect by Provider

Provider	Average Time for a CA to Connect (Seconds)
Provider 1	3.7
Provider 2	5.9

4 The NIST Speech Recognition Scoring Toolkit (SCTK) can be found at http://www.nist.gov/itl/iad/mig/tools.cfm.

5 The providers are not identified to avoid any appearance of endorsement or partiality toward a particular STT service.

Provider	Average Time for a CA to Connect (Seconds)
Provider 3	3.6
Provider 4	3.9
STT-1	N/A
STT-2	N/A

3.1.2 Transcription Accuracy

Table 3 summarizes the average accuracy per provider across audio samples, excluding usability test calls with added noise. Appendix B includes average accuracy data broken out by scenario (conversation). The STT-1 and STT-2 providers are both fully automated speech recognition engines. STT-2 was not tested against all audio samples and is not included in Table 2. Appendix B shows the individual audio sample results for STT-2.

Table 2. Average Accuracy by Provider

Provider	Average Accuracy (Percentage)
Provider 1	88.3
Provider 2	84.5
Provider 3	82.8
Provider 4	88.7
STT-1	83.0

Provider 1 and Provider 4, while generally having longer delay, also generally had higher accuracy than Provider 2 and Provider 3. Interviews with providers indicated that Provider 1 emphasizes higher accuracy, while Provider 2 and Provider 3 emphasize short delay. Survey respondents and usability study participants indicated that both accuracy and delay were areas where improvement was needed.

For all but one test call type, the average accuracy for STT-1, was higher than at least one IP CTS provider. For two audio samples (Pizza – native English Speaker and IRS IVR), STT-1 achieved higher accuracy than any IP CTS provider. This suggests that transcription using automated STT services may be appropriate for some classes of telephony transcription needs.

MITRE observed that faster speech, background noise, more complex speech, computer-generated voices, and non-native English speakers all have a negative impact on accuracy. One script (Pizza) was executed with both a native English speaker and a non-native English speaker. For all providers and SSTs, the average accuracy for the non-native speaker sample was lower than for the native speaker.

Note**:** The Accuracy metric does not account for use of punctuation. Punctuation has an impact on ease of reading not reflected in this metric. Provider 3 transcripts use punctuation, which helps make reading transcripts easier, while other providers and STT-1 do not.

3.1.3 Transcription Delay

Table 3 summarizes the average delay per provider across all test calls, excluding usability test calls with added noise. There is considerable variance in delay between providers, with Provider 1 and Provider 4 generally having higher delay and move variability of delay based on call characteristics. As shown in Appendix B.4, *Conversation #4 – Ordering a Pizza*, all providers maintain relatively low captioning delay during some types of calls.

For all test calls, the delay for providers using Provider 1 and Provider 4 devices tends to increase as each call progresses, only decreasing when there is a pause in the conversation. Both Provider 1 and Provider 4 provide their users with the ability to “clear the buffer” during a call by pressing any button on the dial pad. Clearing the buffer discards the untranscribed audio and “resets” the delay to values typically seen at the beginning of calls.

Both automated STTs tested had low and consistent delay. The STT-2 delay was consistently in the one (1)- to three (3)-second range across all calls. The STT-1 delay was in the one (1) second range for the majority of test samples. As shown in Table 3, the STT captioning delay was significantly lower than the captioning delay for IP CTS providers. This is expected because the IP CTS providers use a person to re-voice audio into a speech to text engine.

MITRE varied noise levels during the usability test calls (conversations #5 through #7). The expected result was that increased noise would result in increased delay. This was not the case. There was no statistically significant correlation between delay and noise level, although accuracy was impacted.

Table 3. Average Transcription Delay by Provider

Provider	Average Transcription Delay (Seconds)
Provider 1	15.8
Provider 2	7.3
Provider 3	4.1
Provider 4	14.6
STT-1	2.2
STT-2	2.1

3.2 Usability Scores

This subsection presents a summary of usability scores based on 20 participants in the baseline usability assessment. Table 4 illustrates the usability scores for each provider. Scores are based on a scale ranging from 1 – 7, where 1 equals Strongly Disagree, 7 equals Strongly Agree, and 4 equals Neither Agree nor Disagree. After completing a call, participants rated the usability of the captions and their overall satisfaction with the call. Higher scores indicate a higher degree of usability.

The results indicate that users generally do not believe that IP CTS transcripts have a high degree of usability. Based on participant feedback, Provider 2 scored the lowest in all usability metrics and Provider 1 scored the highest. Despite Provider 1 having the highest usability ratings, most scores fell just above or below "neither agree nor disagree" (score of 4).

Table 4. Average Caption Usability Scores between Providers

Attribute	Provider 1	Provider 2	Provider 3	Provider 4
Effectiveness	4.73 (SD 2.1)	3.50 (SD 1.9)	3.93 (SD 1.9)	4.60 (SD 1.8)
Efficiency **	4.53 (SD 2.2)	2.56 (SD 1.8)	3.80 (SD 1.6)	4.20 (SD 2.2)
Satisfaction *	4.67 (SD 2.2)	2.94 (SD 1.8)	3.27 (SD 1.6)	4.07 (SD 2.1)
Usability *	4.64 (SD 2.1)	3.00 (SD 1.7)	3.67 (SD 1.5)	4.29 (SD 1.8)

Statistical results present at the ** $p < .05$ and * $p < .07$ significance level.

During usability testing, MITRE observed preferences and behaviors that are not directly reflected in the usability scores in Table 5. These observations include:

- Participants prefer devices with a speaker phone or dual audio option (headphones).
- Participants need on-screen indication(s) that indicate when a dial tone is present and when there is silence on the line, and that provide alternative methods aside from sound, such as haptic feedback, to indicate button selection (tactile feedback).
- Punctuation in captions improves ease of readability and comprehension.

3.3 Survey Results and Findings Detail

MITRE conducted a survey of IP CTS users and controlled usability testing for a limited number of users. MITRE received 540 survey responses, and performed usability testing with 20 individuals.

The majority of survey respondents (88%) indicated that they have at least one year of experience with IP CTS calls. Based on the recent, rapid growth of IP CTS usage MITRE expected a higher percentage of respondents indicating less than one year of usage. Nearly half (44%) of survey respondents reported having profound hearing loss (90 dB or more). Individuals with both profound and lesser hearing loss indicated a need for captioning to make successful calls. A majority of survey respondents (83%) also reported using IP CTS for calls at least several times per week.

All of the controlled usability study participants had received a hearing test from an audiologist or other hearing health care professional, and all participants use hearing devices. Regardless of hearing loss severity, participants used the IP CTS captions during calls made.

Most survey respondents (70%) indicated that they were satisfied or very satisfied communicating with another person when using captions on a call. This result differed from data in the controlled usability assessment, where participants were typically indifferent on their level of overall satisfaction with IP CTS. Potential reasons for this include:

- Providers referenced for this survey may be different from the providers in the controlled usability assessment.
- Scenarios used by survey respondents for their response may be different from the scenarios used in the controlled assessment.
- The controlled usability assessment employed “difficult” scenarios on purpose, which may be reflected in the satisfaction rating.

3.3.1 Accuracy and Delay

MITRE assessed accuracy and delay in both the user survey and usability testing. Survey respondents and usability test subjects indicated that the time it takes for captioning to start was generally acceptable.

Controlled Usability Study participants indicated that they prefer captions that appear quickly (caption delay) and are complete as opposed to accurate transcription. These participants also reported that lack of punctuation greatly affects readability. Provider 3 includes punctuation; other IP CTS providers do not.

Both survey respondents and usability test subjects indicated that caption delay and accuracy are issues. Users want to see improvements in these two areas. A majority do not believe that IP CTS transcripts have a high degree of usability.

3.3.2 Device Characteristics

Almost all survey respondents knew how to turn off the captions (90%). During the controlled usability assessment, MITRE observed that participants were generally familiar or quickly learned how to turn captions on or off for calls.

Survey and Usability study participants identified several characteristics of IP CTS devices where improvements could be made. These include:

- Hard of hearing users benefit from bi-aural (both ears) audio. All Usability test participants indicated a desire to have speakerphone capability.
- Hard of hearing users may not identify audio cues, such as the sound of a dial tone, silence on the line, or the sound of a button being pressed. Usability test participants indicated a desire to have visual or tactile feedback for these types of events.

4. Recommendations

MITRE will continue usability and device testing of IP CTS devices, and will provide further recommendations as additional testing and research phases are completed. Based on the initial testing and findings to date, MITRE recommends that the FCC:

1. Research the possibility of using fully automated Speech to Text services in place of existing IP CTS services. Continue usability testing to determine if automated STT system prototypes can provide (a) similar levels of usability as experienced today by IP CTS users, (b) an overall satisfactory calling experiences similar to what is currently available, and (c) are viable options/alternatives to IP CTS services.
2. Continue working with the hard of hearing community to identify the key measures and metrics for telephony captioning comprehension and usability.
3. Establish an independent quality assurance group to test transcription quality.
4. Request that IP CTS providers include speakerphone and headphone capabilities in devices.
5. Request that IP CTS providers include visual or tactile feedback for all functions that currently use audio feedback (e.g., dial tone or silence indicators, button pressed indicators).

Appendix A. Test Call Transcripts

This appendix provides descriptions and reference transcripts for each test audio sample for the seven (7) conversations.

A.1 Conversation #1: FCC IVR

The first audio sample tested was from a recorded conversation using the FCC Interactive Voice Response (IVR) system reachable at "1-800 call FCC" (1-800-225-5322). The conversation was roughly 2 minutes in duration. IVR systems employ professional voice talents who have clear, understandable speech. Typically, IVR systems have minimal background noise.

Conversation #1 was rated as grade 6.4 based on the Flesch-Kincaid reading ease formula.[6] MITRE calculated the caption delay using the bracketed (and bolded) words in the following transcription of the call:

> Thank you for calling the FCC's consumer **[center]**. Mobile device theft is on the rise in the US and abroad. Be smart about protecting your **[device]**. Use password protection, do not leave your device out in public, and know your device's unique identifying number. Find additional information at fcc.gov/smartdevice. You may have experienced caller id spoofing if you have received a call and you believe the caller has deliberately falsified the telephone number showing on your caller id to disguise their **[identity]**. If so please press **[8]**.
>
> Our hours of operation are from 8 am to 5:30 pm eastern time. To continue in English press 1.
>
> *Note*: MITRE pressed 1
>
> Press 1 if you want to report a tower light outage or safety of life issue involving interference. Press 2 if your call concerns an application for a license or the auctioning of **[frequencies]**. Press 3 if you are a member of the news media. Press 4 if you have a general question or need assistance ordering forms or filing documents. Press 5 if you would like to file a complaint or would like to check on the status of an existing complaint. Press 9 to repeat these **[options]**.
>
> *Note*: MITRE pressed 4
>
> Press 1 if your call concerns any aspect of **[DTV]**. Press 2 if your call concerns debt collection. Press 3 if your call concerns unwanted calls and faxes or to file a complaint. Press 4 if your call concerns obscene or indecent programming, flaming, or non telephone related charges. Press 5 if your call concerns interference to home entertainment equipment. Press 6 if you wish to order forms, documents, or public notices. Press 7 to return to the previous menu. Press 8 to hear these options **[again]**.

6 MITRE calculated the reading level using the online calculator at: https://readability-score.com

A.2 Conversation #2: Paula Thrasher Discussing Road to DevOps Career

This audio sample is an excerpt from Episode 60 of a podcast called "DevOps Café" by John Willis and Damon Edwards (http://devopscafe.org/show/2015/5/15/devops-cafe-episode-60-paula-thrasher.html). The guest on the show, Paula Thrasher, provides an overview of how she came to be involved in Development and Operations (DevOps). The excerpt lasts roughly 90 seconds. It begins about 2.5 minutes into the podcast and demonstrates faster speech, a more casual style of speaking, and more technical content than the FCC IVR Conversation #1.

Conversation #2 was rated as grade 10.5 based on the Flesch-Kincaid reading ease formula via the online calculator used for Conversation #1. MITRE calculated the caption delay using the bracketed and bolded words in the following transcript of the conversation:

> Sure I mean I guess, uh, I always have to sort of start off, how I got into this field is, uh, when I was a kid, that's probably everybody's **[story]**. So, how I learned to code was, uh, when I was a kid there was this thing called the Atari and it had two cartridges. And one cartridge you could put in Basic and the other one was **[blank]**. Uh, and I had some pretty awesome people that taught me how to program Basic, uh, and I wrote my own video games for my Atari which were pretty **[lame]**. They were completely lame video games, but I wrote them and that was what counted, so that's how I got the, uh, the hook into computers and then somewhere along the way I decided you could actually make a living doing **[that]**. Um, not video games unfortunately, um, that didn't work out so well. But uh, um, I worked for an early startup in the late 90s, uh, while I was still in high school actually, uh, which was a great experience and then I went on, uh, to kind of get into the software world. And somewhere along the way, um, a couple years back actually, I was doing my software thing and somebody said, hey we want you to be the IT **[director]**. And I was like I don't know if you want me to do that because I'm a software person and I don't really run servers. You probably want somebody **[else]**. And they asked a couple more times and I finally agreed to do it and actually ended up running, um, IT operations for a small business, about, um, 700 **[folks]**. Um, well I guess small is relative but, um, in, in the federal environment that's considered small. Um, and I was the IT director, sort of a CIO like role over there, um, running our IT operations, and that was I think probably my, like, DevOps awakening **[moment]**.

A.3 Conversation #3: Ms. Jackson

The third audio sample tested was from a conversation in the Virginia Standards of Learning sample test.[7] To mimic a back and forth conversation, a total of 80 seconds of "dead air" was inserted into the 185-second conversation.

Conversation #3 was rated as grade 6.4 based on the Flesch-Kincaid reading ease formula.[8] MITRE calculated the caption delay using the bracketed (and bolded) words in the following transcription of the call:

---

7 The sample test can be found at: http://www.doe.virginia.gov/testing/sol/released_tests/2010/test10_reading8.pdf

8 MITRE calculated the reading level using the online calculator at: https://readability-score.com

In last week's PTA bulletin, I read that the school is still seeking a speaker for this year's eighth-grade awards ceremony. As an eighth-grade student at Eastwood Middle School, I would like you to please consider asking Ms. Ellen Jackson to give this important **[address]**. I realize that the PTA usually asks a local politician or successful business owner to give this inspirational talk, but I think Ms. Jackson would be better than any of our past **[speakers]**.

Ms. Jackson has many fine qualities that make her an excellent choice to speak at the ceremony. She not only meets the requirement of being a former Eastwood Middle School student, but she is also a longtime teacher at our **[school]**. In fact, Ms. Jackson has been associated with this school for much of her life. During that time she has learned many valuable lessons that she passes on to students whenever she **[can]**.

Eastwood Middle School has many fine teachers, and Ms. Jackson is one of the best. Although she teaches English, she is a genius at social studies and math too. Her tutoring sessions are not limited to assignments she has given in her own classes. Ms. Jackson will help with any assignment for any subject. She has, however, one firm rule: when she helps, students must **[work]**. In other words, Ms. Jackson does not simply supply the answers; she teaches students how to find the answers for themselves. She can make difficult concepts seem easy. She patiently explains complicated formulas or confusing procedures one step at a time. Ms. Jackson's homework sessions last as long as necessary. She never ends a session until all students have been **[helped]**.

Ms. Jackson is more than just a great teacher, though. She supports the students of Eastwood Middle School in everything they do. She attends band, orchestra, and choir concerts, and she can be found cheering the Knights to victory at all the school's athletic **[events]**. When Ms. Jackson is absent from one event, it is because she is attending another one. Ms. Jackson also volunteers to chaperone school field trips and outings, including those held during **[summer]** break.

A.4 Conversation #4: Ordering a Pizza

The fourth audio sample tested was from a recorded conversation using a script for ordering pizza. The script includes customer and employee roles, but the only recorded portion is the employee role. This scenario imitated the experience of an IP CTS customer ordering a pizza. The recorded audio was roughly 1 minute in duration and included pauses between the employee's sentences while the customer is speaking.

Conversation #4 was rated as grade 2.5 based on the Flesch-Kincaid reading ease formula. In the following transcription, the lines in italics are *not* included in the recorded audio that was placed under test with the IP CTS devices. They are included here to provide context for the rest of the conversation. MITRE calculated the caption delay using the bracketed (and bolded) words.

Hello. Can I take your **[order]**?

Yes. I'd like a large pepperoni pizza with mushrooms and green peppers.

Would you like anything **[else]**?

Well, wait. Uh, can I make that a half-and-half pizza?

Sure. What would you like on each **[half]**?

I'll have pepperoni and mushrooms on one half and green peppers and Italian sausage on the other.

Okay. Your total comes to fifteen nineteen, which includes tax. And could I have your name **[please]**?

Jay Han.

And your address and telephone **[number]**?

It's 1340 South 16 East, and the phone number is 340-1870.

Okay. Let me repeat your order. A large half-and-half pizza. One half with pepperoni and mushrooms and the other half with Italian sausage and green **[peppers]**. Jay "Han," at 1340 South 16 East, 340-1817. Is that **[correct]**?

Everything except for the phone number. It's 1870, not 1817.

Alright. Thank you for your order. It should arrive at your doorstep in 30 minutes or less, or you'll receive a free small **[pizza]** with your next order.

A.5 Conversation #5: IRS General Information

The fifth audio sample tested was a recorded conversation of a script for requesting an update on a tax return status. The IRS scenario was scripted by calling the IRS Tax Help Line for Individuals and transcribing the IVR paths for a specific task.

Conversation #5 was rated as grade 6.3 based on the Flesch-Kincaid reading ease formula. MITRE calculated the caption delay using the bracketed (and bolded) words in the following transcription of the call:

Welcome to the Internal Revenue Service. You can also visit us at www irs gov.

We are experiencing very high call volumes at this time.

Instead of waiting, you can check your refund, pay your tax bill, or set up a payment plan on irs gov.

If you choose to wait, your call will be processed in the order it was received.

For questions about your refund or to check the status of your form 1040x amended tax return, press 1.

For answers about your personal income taxes, press 2.

For answers about your business taxes, press 3.

To hear general prerecorded information and resources available to you about the healthcare law, press 4.

For answers about your personal or business taxes as it relates to healthcare, press 5.

To repeat this menu, press 9.

[MITRE pressed 1]

If you are calling to check the status of your refund, press 1.

If you are calling to check on the status of a 1040x amended tax return, press 2.

[MITRE pressed 1]

To protect your privacy, you will need the social security number, filing status, and refund amount from your tax return.

If you do not have this information, call us again when you have it available or you may visit our website at www.irs.gov and click on where is my refund.

Please be prepared to take note of important information and call back numbers.

Please enter the social security number for which you are calling.

[MITRE entered a fake SSN: 003-45-6789]

Of the following 5 filing statuses, enter the one used on your return for single, press 1.

For married filing a joint return, press2.

For married filing a separate return, press 3.

For head of household, press 4.

For a qualifying widow or widower with a dependent child, press 5.

[MITRE pressed any number between 1 – 5]

Please enter the exact whole dollar amount of the refund shown on your return.

Do not enter the cent amount.

Enter the amount of your refund followed by the pound sign.

[MITRE pressed 2698]

Your refund was sent to your bank on March 19th 2015.

Please check with your bank or tax preparer if your refund has not been credited to your account.

If you have already checked with your bank and would like to speak to a representative now, press 0.

Or you may call 1-800-829-0582 for assistance between the hours of 7 am and 7 pm Monday through Friday.

When asked for an extension, use 462.

Be sure to have a copy of your tax return with you when you call.

Once again, the number is 1-800-829-0582; when asked for an extension, use 462.

If you would like to hear this information again, press 9.

Thank you for calling the Internal Revenue Service.

A.6 Conversation #6: Requesting a Prescription Refill

MITRE tested a sixth audio sample from a recorded conversation of a script for requesting a refill from a pharmacy. Conversation #4 was rated as grade 9.3 based on the Flesch-Kincaid reading ease formula.[9] MITRE calculated the caption delay using the bracketed (and bolded) words in the following transcription of the call:

Welcome to the MITRE pharmacy automated refill center.

To check the status of a prescription refill request, press 1.

To place a prescription refill order, press 2.

To cancel a prescription refill order, press 3.

[MITRE pressed 2]

Please enter the medical record number followed by the pound sign.

[MITRE pressed 123456#]

To refill pantoprazole, press 1.

To refill metroprolol tartrate, press 2.

To refill levothyroxine, press 3.

To refill cephalexin, press 4.

To refill zetia, press 5.

To refill sertraline, press 6.

To refill Dicyclomine, press 7.

To refill warfarin sodium, press 8.

To refill donepezil, press 9.

To hear your options again, please press the star key.

9 MITRE calculated the reading level using the online calculator at: https://readability-score.com

[MITRE pressed 9]

To request another medication, press 1.

To continue with your order, press 2.

[MITRE pressed 2]

To use the address on file, press 3.

To enter a different address, press 4.

[MITRE pressed 1]

To use the credit card on file, press 1.

To enter a different credit card number, press 2.

[MITRE Pressed 1]

To enter a refill request for another medical record number, press 1.

To complete the transaction, press 2.

[MITRE pressed 2]

Your refill order request has been completed.

Your refill of donepezil will ship in one to two business days.

Thank you for calling the MITRE pharmacy automated refill center.

A.7 Conversation #7: Requesting an Account Balance

MITRE tested a seventh sample from a recorded conversation of a script for requesting an account balance from a bank. Conversation #7 was rated as grade 3.2 based on the Flesch-Kincaid reading ease formula. MITRE calculated the caption delay using the bracketed (and bolded) words in the following transcription of the call:

Welcome to bank of MITRE.

Please enter the last 4 digits of your ATM debit card.

You can also enter your telephone access ID or account number.

[MITRE entered 2222]

Thanks.

Now, please enter the PIN you use with this ATM debit card.

[MITRE entered 1234

One moment.

Please hold while I locate your information.

Your checking account balance is $8,966 dollars and twenty-five cents.

To hear your balance again, please press 1.

For your next payment due, please press 2.

[MITRE pressed 2]

Your next payment is due by august 31st, and the amount due is 478 dollars and ten cents.

This reflects the most current information available on your account.

To end this call, simply hang up.

Appendix B. Quality Metric Summary

B.1 Conversation #1 – FCC IVR

The first audio sample tested was from a recorded conversation using the FCC Interactive Voice Response (IVR) system reachable at “1-800 call FCC” (1-800-225-5322). The conversation was roughly 2 minutes in duration. IVR systems employ professional voice talents who have clear, understandable speech. Typically, IVR systems have minimal background noise. Conversation #1 was rated as grade 6.4 based on the Flesch-Kincaid reading ease formula.

Table 5 summarizes the results for test call #1 by alphabetical listing of providers. Appendix A contains a transcript of the test call.

Table 5. Conversation #1 (FCC IVR) Results Summary

Provider	Accuracy (Percentage)	Average Caption Delay (Seconds)	Average Completeness (Percentage)
Provider 1	92.5	13.7	97.9
Provider 2	90.7	5.9	96.8
Provider 3	89.6	3.8	96.3
Provider 4	93.7	17.4	97.9
STT-1	90.5	1.0	98.8
STT-2	75.8	1.9	96.1

B.2 Conversation #2 – Paula Thrasher Discussing Road to DevOps Career

This audio sample is an excerpt from Episode 60 of a podcast called “DevOps Café” by John Willis and Damon Edwards (http://devopscafe.org/show/2015/5/15/devops-cafe-episode-60-paula-thrasher.html). The excerpt lasts roughly 90 seconds. It begins about 2.5 minutes into the podcast and demonstrates faster speech, a more casual style of speaking, and more technical content than the FCC IVR Conversation #1. Conversation #2 was rated as grade 10.5 based on the Flesch-Kincaid reading ease formula via the online calculator used for Conversation #1.

The STT-1 API consistently timed out after 20 seconds of transcription on this sample. The listed results for STT-1 are based only on the first 20 seconds of the script. MITRE will attempt to resolve this issue and repeat this testing for the full audio sample.

Table 6 summarizes the results for test call #2 by provider. The providers are listed from highest to lowest average accuracy. Appendix A contains a transcript of the test call.

Table 6. Conversation #2 (DevOps) Results Summary

Provider	Accuracy (Percentage)	Average Caption Delay (Seconds)	Average Completeness (Percentage)
Provider 1	81.3	30.1	87.8
Provider 2	72.1	8.4	83.7
Provider 3	75.5	4.1	84.5
Provider 4	81.9	34.1	88.7
STT-1	74.6	1.0	94.6
STT-2	74.8	2.1	92.2

B.3 Conversation #3 – Ms. Jackson

The third audio sample tested was from a conversation in the Virginia Standards of Learning sample test,[10] read by a MITRE employee. To mimic a back-and-forth conversation, a total of 80 seconds of "dead air" was inserted into the 185 second conversation. While this conversation contained pauses in the conversation, the rate of speech is higher than the first two samples. Overall, the accuracy for this sample was lower than for the previous two conversations. In this conversation the accuracy of STT-2 was higher than one of the IP CTS providers, while maintaining lower caption delay than any of the IP CTS providers. Conversation #3 was rated as grade 6.4 based on the Flesch-Kincaid reading ease formula.

Table 7 summarizes the results for test call #3 by provider. The providers are listed from highest to lowest average accuracy. Appendix A contains a transcript of the test call.

Table 7. Conversation #3 (Ms. Jackson) Results Summary

Provider	Accuracy (Percentage)	Average Caption Delay (Seconds)	Average Completeness (Percentage)
Provider 1	84.4	26.3	82.6
Provider 2	79.5	10.9	77.7
Provider 3	62.7	5.1	60.9
Provider 4	87.2	17.7	85.4
STT-1	76.1	1.1	95.5
STT-2	66.5	2.2	63.1

10 The sample test can be found at: http://www.doe.virginia.gov/testing/sol/released_tests/2010/test10_reading8.pdf

B.4 Conversation #4 – Ordering a Pizza

MITRE tested a fourth audio sample from a recorded conversation of a script for ordering a pizza. The script included customer and employee roles, but the only recorded portion is the employee role. This scenario imitated the experience of an IP CTS customer ordering a pizza. The recorded audio was roughly one (1) minute in duration and included pauses between the employee's sentences while the customer was speaking. Conversation #4 was rated as grade 2.5 based on the Flesch-Kincaid reading ease formula.

MITRE recorded the Ordering a Pizza script twice—once with a Native English speaker and once with a non-Native English speaker.

Tables 9 and 10 summarize the results for Conversation #4 by provider for the Native and non-Native English speaker recordings. Each table lists the providers from highest to lowest average accuracy. Appendix A contains a transcript of the test call.

Table 8. Conversation #4 (Pizza) Results Summary – Native Speaker

Provider	Accuracy (Percentage)	Average Caption Delay (Seconds)	Average Completeness (Percentage)
Provider 1	96.2	7.1	99.2
Provider 2	95.8	6.2	99.6
Provider 3	94.5	4.3	98.4
Provider 4	95.1	5.4	98.2
STT-1	98.1	1.0	99.1
STT-2	95.7	2.2	98.2

Table 9. Conversation #4 (Pizza) Results Summary – Non-Native Speaker

Provider	Accuracy (Percentage)	Average Caption Delay (Seconds)	Average Completeness (Percentage)
Provider 1	88.4	7.8	95.6
Provider 2	80.6	5.6	88.6
Provider 3	91.3	4.4	94.4
Provider 4	93.7	4.8	96.7
STT-1	88.2	1.0	96.5
STT-2	65.2	2.1	92.5

B.5 Conversation #5 – IRS IVR

MITRE tested a fifth audio sample from a recorded conversation of a script for requesting an update on a tax return status. The IRS scenario was scripted by calling the IRS Tax Help Line for Individuals and transcribing the IVR paths for a specific task.

Conversation #5 was rated as grade 6.3 based on the Flesch-Kincaid reading ease formula. Some usability test calls included white noise. This is indicated as "audio level in the following tables where "1' indicates no white noise, "2" indicates minimal added white noise, and "3" indicates moderate added white noise. For usability study calls, MITRE added a calculation for reading ease. Tables 11, 12, and 13 summarize the results for Conversation #5 by provider for each audio level. Each table lists the providers from highest to lowest average accuracy. Appendix A contains a transcript of the test call.

Table 10. Conversation #5 (IRS) Results Summary – Audio Level 1

Provider	Accuracy (Percentage)	Average Caption Delay (Seconds)	Average Completeness (Percentage)
Provider 1	92.9	8.6	98.0
Provider 2	94.1	9.2	96.9
Provider 3	90.1	3.5	94.6
Provider 4	78.3	10.5	83.7
STT-1	94.4	1.3	99.3

Table 11. Conversation #5 (IRS) Results Summary – Audio Level 2

Provider	Accuracy (Percentage)	Average Caption Delay (Seconds)	Average Completeness (Percentage)
Provider 1	87.4	10.2	91.8
Provider 2	87.8	6.0	95.0
Provider 3	86.9	4.3	95.3
Provider 4	84.0	4.7	86.2
STT-1	88.9	1.0	98.8

Table 12. Conversation #5 (IRS) Results Summary – Audio Level 3

Provider	Accuracy (Percentage)	Average Caption Delay (Seconds)	Average Completeness (Percentage)
Provider 2	87.7	9.8	94.4
Provider 3	80.7	3.5	88.6
Provider 4	79.8	9.2	85.8
STT-1	85.1	1.0	98.6

B.6 Conversation #6 – Pharmacy

MITRE tested a sixth audio sample from a recorded conversation of a script for requesting a refill from a pharmacy. Conversation #4 was rated as grade 9.3 based on the Flesch-Kincaid reading ease formula.[11] Tables 14, 15, and 16 summarize the results for Conversation #6 by provider for each audio level. Each table lists the providers from highest to lowest average accuracy Appendix A contains a transcript of the test call.

Table 13. Conversation #6 (Pharmacy) Results Summary – Audio Level 1

Provider	Accuracy (Percentage)	Average Caption Delay (Seconds)	Average Completeness (Percentage)
Provider 1	86.3	15.6	5.7
Provider 2	81.6	4.6	89.7
Provider 3	80.4	2.0	96.1
STT-1	68.6	2.6	94.4

Table 14. Conversation #6 (Pharmacy) Results Summary – Audio Level 2

Provider	Accuracy (Percentage)	Average Caption Delay (Seconds)	Average Completeness (Percentage)
Provider 1	73.3	10.4	82.5
Provider 2	67.6	9.2	78.0
Provider 3	69.2	2.8	93.7
Provider 4	79.5	8.1	93.6
STT-1	77.8	1.5	96.7

11 MITRE calculated the reading level using the online calculator at: https://readability-score.com

Table 15. Conversation #6 (Pharmacy) Results Summary – Audio Level 3

Provider	Accuracy (Percentage)	Average Caption Delay (Seconds)	Average Completeness (Percentage)
Provider 1	74.3	6.0	82.7
Provider 2	76.3	7.3	90.1
Provider 3	59.8	2.5	73.9
Provider 4	73.4	3.9	88.7
STT-1	68.0	2.0	96.9

B.7 Conversation #7 – Bank

MITRE tested a seventh sample from a recorded conversation of a script for requesting an account balance from a bank. Conversation #7 was rated as grade 3.2 based on the Flesch-Kincaid reading ease formula. Tables 17, 18, and 19 summarize the results for Conversation #7 by provider for each audio level. Each table lists the providers from highest to lowest average accuracy. Appendix A contains a transcript of the test call.

Table 16. Conversation #7 (Bank) Results Summary – Audio Level 1

Provider	Accuracy (Percentage)	Average Caption Delay (Seconds)	Average Completeness (Percentage)
Provider 1	80.8	5.1	61.9
Provider 2	88.0	6.9	96.1
Provider 3	76.8	3.5	89.0
Provider 4	60.8	6.1	68.5
STT-1	67.5	1.2	88.1

Table 17. Conversation #7 (Bank) Results Summary – Audio Level 2

Provider	Accuracy (Percentage)	Average Caption Delay (Seconds)	Average Completeness (Percentage)
Provider 1	88.7	4.4	92.5
Provider 2	81.2	8.6	91.2
Provider 3	68.9	22.0	85.2
Provider 4	88.5	4.0	91.5
STT-1	74.6	1.0	91.1

Table 18. Conversation #7 (Bank) Results Summary – Audio Level 3

Provider	Accuracy (Percentage)	Average Caption Delay (Seconds)	Average Completeness (Percentage)
Provider 1	79.3	4.8	87.5
Provider 2	69.9	6.5	81.0
Provider 3	83.4	4.1	93.4
Provider 4	82.4	3.0	88.4
STT-1	63.6	1.0	87.4

Appendix C. IVR Scenario Transcripts

C.1 IP CTS Bank Scenario Transcript

Table 19. Transcript for IP CTS Bank Scenario

Component	Content
Scenario Name	Bank Scenario
Objective	Participant will contact their bank for a current balance and amount of next payment
Description	The person will enter IVR instructions on the phone's key-pad or respond verbally. Keywords used for delay calculation in **bold**.
Provide	Last four digits of ATM card: 1234
Scenario Script / Steps	**Step 1:** [Participant dials bank phone number. Bank IVR answers.] **Bank IVR:** Welcome to Bank of **MITRE**. Please enter the last four digits of your ATM debit card. You can also enter your telephone access ID or account number. [Participant enters "1234"] **Bank IVR:** Thanks. Now, please **enter** the pin you use with this ATM debit card. [Participant enters 2222] **Bank IVR:** One moment. Please hold while I **locate** your information. [Pause for three seconds] **Bank IVR:** Your checking account **balance** is $8,966 dollars and twenty-five cents. To hear your balance again, please press 1. For your next payment due, please press 2. [Participant presses 2] **Bank IVR:** Your next **payment** is due by August 31 and the amount due is $478 dollars and ten cents. This reflects the most current information available on your account. To end this call simply **hang up**. [Participant hangs up]

C.2 IP CTS Pharmacy Scenario Transcript

Table 20. Transcript for IP CTS Pharmacy Scenario

Component	Content
Scenario Name	Pharmacy Scenario
Objective	For the participant to submit a prescription refill request through an automated system
Description	Order medication Keywords used for delay calculation in **bold**.
Provide	Script
Scenario Script / Steps	[Participant dials pharmacy phone number. Pharmacy IVR answers.] **Pharmacy IVR:** Welcome to the MITRE Pharmacy Automated Refill Center. To check the status of a **prescription** refill request, press 1. To place a prescription refill order, press 2. To cancel a prescription refill order, press 3. [Participant presses 2] **Pharmacy IVR:** Please enter the medical record number, followed by the **pound** sign. [Participant enters 123456#] **Pharmacy IVR:** To refill Pantoprazole, press 1. To refill Metroprolol Tartrate, press 2. To refill **Levothyroxine**, press 3. To refill Cephalexin, press 4. To refill Zetia, press 5. To refill Sertraline, press 6. To refill Dicyclomine, press 7. To refill Warfarin **Sodium**, press 8. To refill Donepezil, press 9. To hear your options again, please press the star key. [Participant enters 9] **Pharmacy IVR:** To request another medication, press 1. To **continue** with your order, press 2. [Participant presses 2] **Pharmacy IVR:** To use the address on file, press 1. To enter a **different** address, press 2. [Participant presses 1]

Component	Content
Scenario Script / Steps	**Pharmacy IVR:** To use the credit card on file, press 1. To enter a different **credit** card number, press 2. [Participant presses 1] **Pharmacy IVR:** To enter a refill request for another **medical** record number, press 1. To complete the transaction, press 2. [Participant presses 2] **Pharmacy IVR:** Your refill order request has been **completed**. Your refill of Donepezil will ship in one to two business days. Thank you for calling the MITRE Pharmacy Automated **Refill** Center.

C.3 IP CTS IRS Scenario

Table 21. Transcript for IP CTS IRS Scenario

Component	Content
Scenario Name	IP CTS IRS Scenario
Objective	Participant will contact IRS for update on tax return
Description	The person will enter IVR instructions on the phone's key-pad or respond verbally. Keywords used for delay calculation in **bold**.
Provide	SSN#: 003-45-6789 Filing status: Joint Refund Amount: $2698 (2015 Average as of 5/5/15)
Scenario Script / Steps	[Participant dials phone number. IVR answers. Participant selects # for IRS scenario] **IRS IVR:** Welcome to the Internal **Revenue** Service, you can also visit us at W. W. W. dot I. R. S. dot com. We are experiencing very high call volumes at this time. Instead of waiting, you can check your **refund**, pay your tax bill, or setup a payment plan on I. R. S. dot gov. If you choose to wait, your call will be **processed** in the order it was received. For questions about your refund or to check the status of your form 10 40 X **amended** tax return, press 1. For answers about your **personal** income taxes, press 2. For answers about your business taxes, press 3. To hear general prerecorded information and resources available to you about the **health** care law, press 4. For answers about your personal or business taxes, as it **relates** to health care, press 5.

Component	Content
Scenario Script / Steps	To repeat this menu, press 9. [Press 1] [if 9 pressed, repeat menu] [if any other number pressed say “I’m sorry, that option is not available at this time” and repeat menu] If you are calling to **check** the status of your refund, press 1. If you are calling to check on the status of a 10 40 X amended tax return, press 2. [Press 1] [if any other number pressed say “I’m sorry, that option is not available at this time” and repeat menu] To protect your **privacy** you will need the social security number, filing status, and refund amount from your tax return. If you do not have this information, call us again when you have it available. Or you may visit our website at W. W. W. dot. I. R. S. dot gov. and **click** on “where’s my refund.” Please be prepared to take note of important information and call back numbers. Please enter the **social** security number for which you are calling. [Enter fake SSN] Of the following five filing statuses, enter the one used on your return. For Single, press **1**. For Married filing a **joint** return, press 2. For Married filing a separate return, press 3. For Head of household, press 4. For a Qualifying widow or **widower** with a dependent child, press 5. [Press any number between 1-5] Please enter the exact whole dollar amount of the refund shown on your return. Do not enter the cent **amount**. Enter the amount of your refund, followed by the pound sign. [Refund amount & #] -------More on next page. Your refund was sent to your bank on March 19th, 2015 please check with your bank or tax preparer if your refund has not been **credited** to your account. If you have already checked with your bank and would like to speak to a representative, now, press 0. Or you may call 1. 800. 8. 2. 9. 0-5. 8-2. For **assistance** between the hours of 7 A. M. and 7 P. M. Monday through Friday. When asked for an **extension** use 4-6-2. Be sure to have a copy of your tax return with you, when you call. Once Again, the number is 1. 800. 8. 2. 9. 0-5. 8-2. When asked for an extension use 4-6-2. If you would like to **hear** this information again press 9. [If 9 pressed, repeat refund message.] Thank you for calling the Internal Revenue Service. [Disconnect or Participant Hangs up]

Appendix D. Usability Questionnaire

Please circle a rating for the audio quality, captioning quality and usability of the IP CTS device for each of the three statements:

1) Throughout the phone call, I was able to maintain full context and meaning of the:

Audio (the quality of the sound)

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neither disagree or agree	Somewhat agree	Agree	Strongly agree

Captioned text of the call on the screen

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neither disagree or agree	Somewhat agree	Agree	Strongly agree

2) I felt comfortable with the speech rate of the phone call:

Audio (the quality of the sound)

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neither disagree or agree	Somewhat agree	Agree	Strongly agree

Captioned text of the call on the screen

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neither disagree or agree	Somewhat agree	Agree	Strongly agree

Figure 1. Usability Questionnaire, Page 1

3) I was satisfied with the overall experience with the:

Audio (the quality of the sound)

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neither disagree or agree	Somewhat agree	Agree	Strongly agree

Captioned text of the call on the screen

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neither disagree or agree	Somewhat agree	Agree	Strongly agree

Usability of the IP CTS Device

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neither disagree or agree	Somewhat agree	Agree	Strongly agree

For MITRE staff use:

Participant ID:	
Scenario #:	
Date:	
Session Time:	

Figure 2. Usability Questionnaire, Page 2

Acronyms

ASR	Automatic Speech Recognition
CA	Communication Agent
CAMH	CMS Alliance to Modernize Healthcare
CMS	Centers for Medicare & Medicaid Services
CTS	Caption Telephone Service
dB	Decibels
DUT	Device under Test
FCC	Federal Communications Commission
FFO	Baseline Usability Assessment scenario: Phone a Friend, Family Member, or Public Organization
FFRDC	Federally Funded Research and Development Center
IP	Internet Protocol
IRS	Internal Revenue Service
IVR	Interactive Voice Response
NIST	National Institute of Standards and Technology
PSTN	Public Switched Telephone Network
SCTK	Speech Recognition Scoring Toolkit
SD	Standard Deviation
STT	Speech to Text
TRS	Telecommunications Relay Services
TTS	Text to Speech
UA	Usability Assessment
URL	Universal Resource Locator
VRS	Video Relay Services
WER	Word Error Rate